(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（A）　(11)特許出願公開番号

特開平4－337850

(43)公開日　平成４年(1992)11月25日

(51)Int.Cl.[5]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ　12/00	５１８　Ａ	8944－5Ｂ		

審査請求　未請求　請求項の数10(全 14 頁)

(21)出願番号　特願平3－330889

(22)出願日　平成３年(1991)12月13日

(31)優先権主張番号　６６０７６９

(32)優先日　1991年２月25日

(33)優先権主張国　米国（US）

(71)出願人　390009531
インターナショナル・ビジネス・マシーンズ・コーポレイション
INTERNATIONAL BUSINESS MASCHINES CORPORATION
アメリカ合衆国10504、ニューヨーク州　アーモンク　（番地なし）

(72)発明者　ダニエル、マヌエル、デイアス
アメリカ合衆国ニューヨーク州、マホパック、パイク、プレイス、16

(74)代理人　弁理士　頓宮　孝一　（外５名）

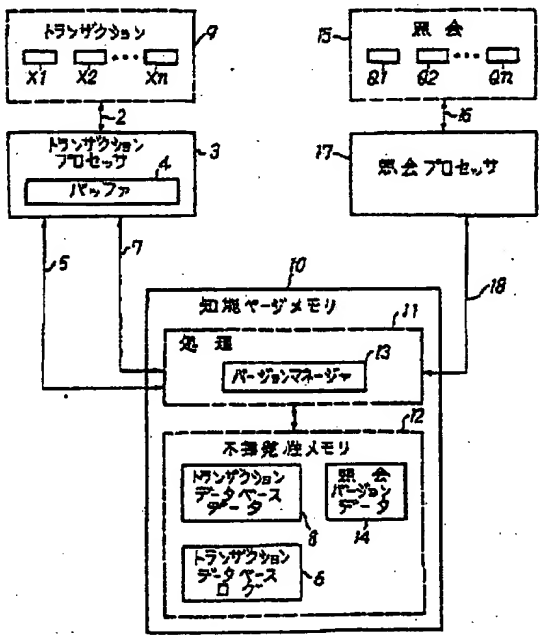
最終頁に続く

(54)【発明の名称】　データベース・トランザクション及び照会処理システム

(57)【要約】

【目的】　知能ページ・メモリを用いて実現される機能的に分離されたトランザクション主体及び照会主体に共有データベースへの同時及び一貫したアクセスを提供する一方でデータの殆どの単一の物理コピーを維持するための方法及び装置を提供する。

【構成】　知能ページ・メモリ（１０）は、共有不揮発性メモリ（１２）を含み、知能バージョニング機構（１３）がトランザクション主体（３，９）及び照会主体（１５，１７）による共有データへの同時アクセスを可能にする。トランザクション主体には、現データが提供され、照会主体には、データの最近の一貫したバージョンが与えられる。最近更新されたページを除く全てのページの単一のコピーが知能ページ・メモリによって維持される。照会及びトランザクション主体は、互いに独立して動作し、別個に最適化される。



1

【特許請求の範囲】

【請求項1】データベースの個々のページの主バージョンに対する不揮発性メモリとして使用され、また最も最近に更新されたデータベース・ページに対するアクセスを必要としない照会によってアクセスされるための前記データベースの少なくとも一つの一貫したスナップ・ショット・バージョンを生成及び維持するための一つの知能ページ・メモリと、前記データベースの前記主バージョン・ページにアクセスし、これを更新するトランザクションプロセッサと、照会を前記データ・ベースの前記少なくとも一つの一貫したスナップ・ショット・バージョンに対して実行するための前記トランザクションプロセッサとは独立した一つの照会プロセッサとを備え、前記知能ページ・メモリが前記データベースの任意のページの前記主バージョンと同一の一つの物理コピー及び前記データベースの前記少なくとも一つのスナップ・ショット・バージョンのみを維持し、前記主バージョン・ページが前記トランザクションプロセッサに前記知能ページ・メモリによって供給され、前記少なくとも一つの一貫した前記データベースのバージョンが前記照会プロセッサに供給されることを特徴とするデータベース・トランザクション及び照会処理システム。

【請求項2】前記トランザクションプロセッサと前記照会プロセッサが異なる物理主体から成ることを特徴とする請求項1に記載のデータベース・トランザクション及び照会処理システム。

【請求項3】前記トランザクションプロセッサと前記照会プロセッサが同一の物理主体上で実行される独立したプロセスであることを特徴とする請求項1に記載のデータベース・トランザクション及び照会処理システム。

【請求項4】前記ページ・メモリが前記データベースの一つ以上の一貫したスナップ・ショット・バージョンを維持することを特徴とする請求項1に記載のデータベース・トランザクション及び照会処理システム。

【請求項5】前記トランザクションプロセッサが前記データベースの幾らかのページの最も最近のバージョンに対する高速アクセス・メモリとして機能するページ・バッファを含み、前記知能ページ・メモリが前記ページ・バッファ内に前記最も最近のバージョンのページの全てのコピーを必ずしも含まないことを特徴とする請求項1記載のデータベース・トランザクション及び照会処理システム。

【請求項6】前記知能ページ・メモリが前記データベースの一貫したスナップ・ショット・バージョンを前記ページ・バッファを前記知能ページ・メモリ内にフラッシングすることによって作成し、前記フラッシング後の前記データベースの前記主バージョンが一つの一貫したスナップ・ショット・バージョンとなることを特徴とする請求項5に記載のデータベース・トランザクション及び照会処理システム。

2

【請求項7】データ・ベース・ログが含まれ、前記知能ページ・メモリが前記主バージョン・ページ及び前記データベース・ログから一貫したスナップ・ショット・バージョンを派生することを特徴とする請求項1に記載のデータベース・トランザクション及び照会処理システム。

【請求項8】前記データベース・ログが前記知能ページ・メモリ内に格納されることを特徴とする請求項7に記載のデータベース・トランザクション及び照会処理システム。

【請求項9】前記知能ページ・メモリが前記データベースの各々のページの前記主バージョンを格納するための第一のメモリ空間、及び前記データベースの幾らかのページの少なくとも一つのより古いバージョンを格納するための第二のメモリ空間を含むことを特徴とする請求項1に記載のデータベース・トランザクション及び照会処理システム。

【請求項10】データベース・トランザクション及び照会処理システムにおいて、前記システムが：トランザクション処理主体；照会処理主体；及び一つの順番のセットの時間間隔を持つ知能ページ・メモリを含み、現時間を含む時間間隔が現時間間隔であり、前記知能ページ・メモリが複数の論理ページから構成されるデータベースに対する不揮発性メモリとして機能し、各々の前記論理ページが一つあるいは複数の物理ページと関連し、各々の前記論理ページと関連する前記物理ページの一つが前記主物理ページであり、各々の前記物理ページと関連する他の物理ページが補助ページと呼ばれ、各々の補助物理ページが一つのタイム・ステップと関連し、前記トランザクション処理主体と前記照会処理主体が前記知能ページ・メモリへのページ・アクセスを同時に行い、前記トランザクション処理主体のページ要求が読出しあるいは書込み要求であり、前記照会処理主体のページ要求が読出し要求であり；前記知能ページ・メモリがさらに：前記トランザクション処理主体に前記トランザクション処理主体によって要求された論理ページに対応する主物理ページをリターンするための手段；特定の照会に応答して前記照会処理主体に前記特定の照会と関連する時間間隔に対応する補助ページ、あるいは前記特定の照会と関連する時間間隔に対応する補助ページが存在しない場合に前記主物理ページをリターンするための手段；前記トランザクション処理主体が前記関連する論理ページを更新するとき、その論理ページと関連する現時間ステップと対応する補助物理ページが存在しない場合、新たな補助物理ページを作成するための手段；新たな時間間隔を前記新たな時間間隔と関連する主及び補助ページが前記データベースの論理的に一貫したセットのページと対応するように生成するための手段；及び；現時間間隔以外の任意の時間間隔と関連する補助ページの全てを前記時間間隔と関連する全ての照会が完了した後に削除する

3

ための手段を含むことを特徴とするデータベース・トランザクション及び照会処理システム。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、一般的には、同時データベース・トランザクション処理及び照会処理、より具体的には、知能ページ・メモリと呼ばれる方法及び装置を提供することによってデータベース・トランザクション主体と同一のデータにアクセスする照会主体とを機能的に分離することに関する。この知能ページ・メモリは、トランザクション主体に対して一つ、そして照会主体に対して一つの計二つのアクセス経路を提供する。本発明においては、さらに、暗黙バージョニング機構が提供されるが、これは、トランザクション主体に最も最近のデータを提供し、照会主体にデータの最近の一貫したバージョンを提供する一方において、殆どのデータの単一コピーを物理的に保持する。

【０００２】

【従来の技術】近年、大きな設備におけるデータベース・トランザクション処理能力に対する需要が著しく増加している。同時に、新たなデータベース・アプリケーションの大部分は、関係データベース環境（relational database environment）内におけるものであり、この関係データベース環境はまたデータベース上のその場限りの照会をサポートするための理想的な環境でもある。この結果として、その場限りの構造化されない照会の使用が付随的に増加してきており、この傾向は加速されるものと予測される。このために、同一データベースに対して、高容量のトランザクション処理及び構造化されない照会の両方を同時にサポートすることに対する要求が増大している。従って、本発明の主要な目的は、高容量トランザクションと複合照会の両方を、これらの間の干渉が最少に留まるような方法で、しかもデータの殆どの単一コピーを共有して、効率的にサポートするアーキテクチャを設計することにある。

【０００３】典型的には、企業は、それらのデータベースを高容量の比較的単純なトランザクションを通じて保持する。個々のトランザクションは、良く知られた商業活動を表わす（新たな顧客レコードが記録され、勘定の支払いあるいは振替えが記される）。ますます多くの企業がそれらのオン・ライン・データに対してより多くのその場限りの構造化されない照会をランすることに関心を持つようになってきている。これは、これらより複雑な照会をＳＱＬで書くことができることから刺激されたものである。典型的なアプリケーションとしては、新たな市場機会のテスト、判定のサポート、歴史的傾向の検出、利潤分析等が含まれる。これら構造化されない照会は、以下の特徴を持つ。・これらは、計画されたものでなく、頻繁には使用されない；各々の照会に対する性能調整は現実的でない。・これらは運営事業データを変更

4

しない。・これらは価値を失うことなく幾分古いデータベース・データに対して実行できる。・これらは大きな量のデータ走査及び処理を要求する；従って、通常のトランザクションと比較して長い実行時間を持つ。

【０００４】チャン（Chan），Ａ．、フォックス（Fox），Ｓ．、リン（Lin）Ｗ．Ｔ．、ノリ（Nori），Ａ．及びリーズ（Ries），Ｄ．らによって、Proc.ACM SIGMODConf.，１９８２、ページ１８４－１９１に掲載の論文『総合同時制御及び回復体系の実現（The Implementation of an Integrated Concurrency Controland Recovery Scheme）』において、バーショニング体系について説明されている。この体系においては、ページの異なるバージョンが連鎖され、再び個々のバージョンがそれを生成したＩＤトランザクションによって識別される。各々の照会は、それと関連して完結されたトランザクションのリスト（Completed Transaction List、ＣＴＬ）のコピーを持つが、これは、事実上、その初期の時のものである。照会アクセスは、この物理ページの連鎖を照会ＣＴＬ内の一つのバージョンが検出するまで追跡することによってアクセスされる。第一に、これは、完結されたトランザクションの情報が照会プロセッサに供給できることを必要とし、ここでも、トランザクション処理と照会処理を機能的に分離することを機能的に阻止する。第二に、ページの追跡には複数のＩ／Ｏが要求される。第三に、この体系においては、ページはディスクにトランザクションをコミットすることによって入れられなければならない。最後に、この体系は、トランザクションに対するページ・レベル・ロッキングのみをサポートする。この体系は、チャン（Chan），Ａ．及びグレイ（Gray），Ｒによって、IEEE Trans. Software Engrg.,，Ｖｏｌ．ＳＥ－１１、Ｎｏ．２、１９８５年２月号に掲載の論文『分散読出し専用トランザクションの実現（Implementing Distributed Read-Only Transactions）』において、ＣＴＬがサイト間で送られ、これらがＣＴＬの新たなバージョンを生成するために併合される複雑な体系を使用することによって分散環境に一般化される。

【０００５】ロビンソン（Robinson），Ｊ．、トーマシアン（Thomasian），Ａ．及びユー（Yu），Ｐ．は、ＩＢＭ技術開示誌報、Ｖｏｌ．３１、Ｎｏ．１、１９８８年６月号、ページ１８０－１８５に掲載の論文『読出し専用照会及びオンライン更新トランザクションによってアクセスされる関係データベース内のロック競合の排除（Elimination of Lock Contention in Relational Databases Accessed byRead-Only Queries and On-Line Update Transactions）』において、両者がデータに共通の同時コントローラの場所で要求することによってアクセスする照会及びトランザクションに対する明示的ページ・バージョニング法について説明する。この体系は、どのページがトランザクション及び照会によってロック

5

されたかの知識を必要とし、そして、ロック競合が検出されると、アクセスのために照会に対する一つのバージョンが生成される。照会に対しては、これは、進行中の照会の状態アレイを保持し、トランザクションが結果として衝突を起こすロック要求を行なったときこれらアレイをチェックすることによって行なわれる。この体系はまたトランザクションによってコミットされた更新されたページが照会によって直ちにアクセスできることを要求する。本質的に、この体系は、照会及びトランザクションによって行なわれてロックについて互いに知ることができるように、照会及びトランザクションが単一のＤＢＭＳ（共通の同時実行制御マネージャ及びバッファ・マネージャ）の下で実行されることを要求する。この開示は、もはや必要でなくなったバージョン化されたページを除去するためにどのように廃物の収集が行なわれるかについては記述しない。

【０００６】

【発明が解決しようとする課題】関係データベース照会がより複雑になるにつれて、同一照会に関して協力する多数のプロセッサを活用するパラレル照会内処理（parallel intra-queryprocessing）が照会応答時間を向上させるため、及び増分的な拡張を行なうための重要な手段となってきている。他方、トランザクション処理は、殆どの部分がトランザクション内パラレル処理（intra-transaction parallelism）に修正することができず、サブ秒の応答時間を持つ多数の同時トランザクションをサポートすることを要求する。トランザクション速度の増加にともなってロック保持時間を短縮してデータ競合を低減することが非常に重要になる。これは、共有バッファを持つ大きなプロセッサを支持する。従って、本発明の主要な目的は、データにアクセスするための二つの経路を持つ論理データベースを提供することにある。一つは、データベース・トランザクションのために使用され、もう一つは、その場限りの照会のために使用される。これは、トランザクション及び照会処理システムを独立して最適化し、他方において、同一データへのアクセスを可能にする。例えば、更新及びトランザクション・トラヒックは、密に結合されたメモリ構成内での多数のプロセッサの性能を活用でき、他方、同一データに対する複合照会は、粗く結合されたマイクロプロセッサ上のパラレル・データベース・ソフトウエアによって扱うことができる。

【０００７】上の特性を持つトラヒック及び照会をサポートする環境内において、本発明のさらに別の目的は以下の通りである。・ディスク及びディクス・コンピュータは大きなデータベースが存在する場合総コストの大きな部分を占める。従って、組合わされたトランザクション及び照会処理のためのディクス空間は最小限にされるべきである。従って、本発明の一つの目的は、オンライン・データがトランザクション及び照会によって共有さ

6

れるようにすることにある。・複合照会は、しばしば、スループットが重要である”構造化されたトランザクション”の応答時間よりもかなり長い実行時間、及びロック保持時間を持つ。従って、本発明のさらにもう一つの目的は、複合照会がトランザクション・トラヒックからのロックを差し控えることなくデータベース・データの一貫してビューを見るようにすることである。・本発明のさらにもう一つの目的は、トランザクション処理データベース・ソフトウエア及び照会処理ＤＢソフトウエアが効果的に分離されるようにすることである（メモリ内の互いのバッファへのアクセス、あるいはロック情報の交換は行なわれない）。一般的に、これは、トランザクション及び照会処理に対するソフトウエアを独立して最適化することを可能にする。

【０００８】

【課題を解決するための手段】本発明の目的及び特徴を示す一つの好ましく、ただし、一例としての実施例によると、共有ディスク・メモリを含む知能ページ・メモリが提供されるデータベース・トランザクション処理及び照会処理を同時に行う新規の方法及び装置が提供される。この知能ページ・メモリは、この共有データへの二つのアクセス経路を提供するが、一つはトランザクション主体によって、もう一つは照会主体によって使用される。この知能ページ・メモリ内においては、暗黙バージョニング機構が共有ディスク・メモリへのトランザクション主体及び照会主体の同時アクセスを可能にする。ここで、トランザクション主体には、現データが提供され、照会主体には、データの最近の一貫したバージョンが提供される。さらに、最近更新されたページを除く全てのページの単一のコピーが知能ページ・メモリによって維持され、照会及びトランザクション主体は、互いに独立して動作する。

【０００９】上に要約された先行技術と本発明との大きな差異は、先行技術においては、照会とトランザクションが互いに機能的に分離されないことである。つまり、先行技術においては、様々なトランザクションの間、及びトランザクションと照会との間の一貫したアクセスを保証するための単一の同時実行制御主体が存在する。これは、照会及びトランザクション処理の独立した実現及び最適化の可能性を排除する。先行技術による方法のもう一方の極端においては、データベースの一つの完全なコピーを作成することによって照会及びトランザクションによってアクセスされるデータが分離される。

【００１０】本発明のこれら及びその他の目的、長所、及び特徴は、以下の説明及び付属の図面から一層明らかになるものである。

【００１１】

【実施例】「先行技術によるトランザクション及び照会システム」本発明の説明を簡素化するために、トランザクション及び照会処理のための先行技術によるシステム

7

の基本要素を図1に簡略的に示す。

【0012】この図面において、ボックス1は、データベースに対して実行されるセットのトランザクションX1、X2、．．．．Xm及び照会Q1、Q2、．．．Qnを示す。我々の環境内においては、十分なスループットを達成するために、多くのトランザクションがデータに対して同時に実行される必要性を持つ。ただし、これらプログラムによって見られるデータのビューは、これらが、シリアルに、アトミック的に、干渉されることなく実行されるようなものである。

【0013】トランザクションプロセッサ3は、トランザクション1によって期待されるデータのビューを満たすために必要とされる同時実行コントロール、ロッキング、データ・アクセス・チェッキング、インデックス管理、緩衝及びデータ保護を提供する。最も以前のシステムにおいては、更新トランザクション及び読出しのみの照会は、分離されることなく、独立して扱われた。トランザクションプロセッサは、典型的には、トランザクションの要求されるスループットをサポートするために調整される。幾つかの同時実行照会もこのトランザクションプロセッサによってサポートされるが、ただし、並行して実行されるより長い読出しのみの照会が存在する状況において、更新トランザクションの高いスループットが要求される場合は、困難が発生する。

【0014】トランザクション1とトランザクションプロセッサ3との間の対話は、トランザクションを実行あるいは中断するとき、データ・ベース・レコード及び情報を読出したり書込んだりするためのデータ・アクセス要求から成る。これは、インターフェース2として示される。

【0015】トランザクションプロセッサ内のワード・プロセッサは、特定のシステム編成、マシーン・パッケージあるいは物理ユニット境界を意味するものではない。トランザクションプロセッサ機能は、単一の物理プロセッサ上で実行することも、一群のプロセッサの多重プロセッサ網上で実行することも、あるいは処理システムを他の機能と共有する要素上で実行することもある。以降、我々は、プロセッサをこのように自由に使用する。

【0016】トランザクションプロセッサの重要な要素は、バッファ4である。これは、トランザクションプロセッサによって管理されるプールの高速アクセス・メモリ（例えば、電子メモリ）である。データベース・データを読出したり、あるいは修正するためには、ページを不揮発性媒体（例えば、磁気ディスク）上に格納されたトランザクション・データベース・データ8のページからこのバッファ内に読み込むことが必要である。更新は、このバッファ内のページに対して行なわれ、これが最終的には、この不揮発性データベース・データ・メモリ内に書き戻される。

8

【0017】トランザクション処理は、また、データベース・データをトランザクション破棄及びシステム故障から保護するために、トランザクション・データベース論理6を使用する。このトランザクション・データベース論理は、不揮発性メモリ（例えば、磁気ディスク）上に格納されなければならない。トランザクションプロセッサが何をそのログ内に書込むべきか、及びこれがバッファ・マネジメントといかにコーディネイトされるべきかを記述する標準用語およびアルゴリズムについては以下の文献において説明されている。つまり、C．モーハン（Mohan）、D．ヘダーレ（D.Haderle）、B．リンドセイ（Lindsay）、H．ピラヘシ（Pirahesh）、P．シュワルツ（Schwarz）らによって、IBMリサーチ・リポートRJ6649、1／23／89に掲載の論文『ARIES：ライト・アヘッド・ロッキングを使用して細かい粒度のロッキング及び部分ロールバックをサポートするトランザクション方法（A Transaction method supporting finegranularity, locking and partial rollbacks using write ahead locking）』、及びR．A．クルーズ（Cruz）によって、IBMシステム・ジャーナル、Vol．23、no2、1984に掲載の論文『IBMデータベース2におけるデータの回復（Data Recovery in IBM Database 2）』において説明されている。

【0018】トランザクション・データベース論理6及びトランザクション・データベース・データ8は、両方とも、不揮発性メモリ内に格納されなければならない。これは、通常、磁気ディスクを使用して供給されるが、ただし、任意の匹敵するメモリ媒体を使用することができる。トランザクションプロセッサは、論理に対してインターフェース5を介して、またデータベース・データに対してインターフェース7を介して、一連のページ読出し及び書込み要求を発行する。これらページ読出し及び書込みコマンドの内容は、トランザクションプロセッサによって定義される。インターフェース5及び7によって要求される不揮発性メモリ・サービスを提供するための基本要件は、一旦書込まれた任意のページの内容が後の任意の時間において正確に検索できることである。

【0019】「知能ページ・メモリに対するシステム構造；方法及び装置」図2には、本発明のためのシステム構造、つまり、照会及びトランザクションの同時実行を可能とするための知能ページ・メモリ・体系のブロック図が示される。これは、トランザクションプロセッサ3、照会プロセッサ17、及び知能ページ・メモリ10から構成される。このシステムと先行技術によるトランザクション及び照会システムとの間の基本的な差異は、更新トランザクションが照会と分離され、トランザクションプロセッサ及び照会プロセッサによって独立して処理されることである。この知能ページ・メモリは、トランザクション及び照会プロセッサに、重要な性能要件（同時トランザクション及び照会処理）が不揮発性メモ

9

リ要件の観点から最少コストでサポートされるような方法により、データベースの共有物理ページへのアクセスを与える。

【0020】トランザクションプロセッサによってサポートされるセットのトランザクション9は、データベースからの現在のデータへのアクセスを要求する更新トランザクション及び読出しのみのトランザクションのみを含む。これは、図1のトランザクションプロセッサ3によって扱われるトランザクション及び照会1のサブセットX1、X2、．．．Xmによって表わされる。これらトランザクションは、データベースを先行技術と同じビューで見る。つまり、アトミックで、非干渉的な、シリアル化された動作として見る。こうして、これらがトランザクションプロセッサ3と、レコード・アクセス要求及び実行／中止情報の同一のインターフェース2を介して対話する事実が図解される。

【0021】この図面内のトランザクションプロセッサ3及びそのバッファ4の機能は、図1と同一である。これらは、一体となって、トランザクションX1、X2、．．．Xmによって期待されるデータのビューをサポートするための緩衝ロッキング、同時制御、インデックス管理サービスを提供する。トランザクションプロセッサは、その回復ログに、ページ読出し及び書込み要求のストリームをインターフェース5を介して、あたかも回復ログがそれに接続された不揮発性メモリ上に直接に格納されているかのように書込み続ける。

【0022】照会プロセッサ17は、それが一貫するものであると言う前提の下で、完全に現在のものではないデータを受け入れることができる書込みのみの照会をサポートする。セットの照会15が先行技術においてはトランザクションプロセッサによって処理される一例としての照会Q1、Q2、．．．Qnにて示される。照会Q1、Q2、．．．Qnは、これらが独立して処理できるように明確に識別され、トランザクションX1、X2、．．．Xmから分離される。この別個のインターフェースを介して受信される照会は、これらは読出しのみであり、完全に現在のデータである必要がないと言う前提の下で処理される。照会からの要求のみを含むことが知られているレコード読出しアクセスのストリームであるこの別個のインターフェースがインターフェース16として示される。

【0023】照会プロセッサ17は、照会に対するアクセス・チェッキング、インデックス管理、緩衝等を提供することを要求される。照会プロセッサ17のために先行技術においてトランザクションプロセッサ3のために使用されたのと同一のソフトウエア及びハードウエア処理を使用することも可能である。ここで、先行技術においては、照会も処理する。ただし、殆どの現存のデータベース・システムは、十分なトランザクション・スループットがサポートできると言う要件にてドライブされて

10

いるために、照会処理を別個の主体によって処理し、ソフトウエア及びハードウエア処理を再調整あるいは再設計し、読出しのみの照会のみをサポートするようにすることができる。照会のみのシステム内における同時制御、ロッキング及びデータ保護に対する要件は、汎用トランザクション処理のものと比較して軽減されるため、かなりの性能及びコスト上の利益が得られる。

【0024】照会プロセッサ17は、幾らかの内部ページ・バッファを含むが、ただし、バッファ管理体系は、バッファ4において使用されるのと異なるために、バッファはサブ要素として明確には定義されない。

【0025】インターフェース18は、照会プロセッサ17がページ読出し要求のストリームを介してデータベース・データのページを要求することを可能にする。こうして供給されるデータは、最近のデータベース・データの一貫したビューを表わす。

【0026】知能ページ・メモリ10は、新しい概念であり、不揮発性メモリの大きなコストないしトランザクションと照会の分離した処理を（それぞれトランザクション及び照会処理によって使用すための全てのデータベース・データの独立したコピー）を可能にする。

【0027】知能ページ・メモリは、処理部11及び不揮発性メモリ部12を含む。知能ページ・メモリ内の処理は、インターフェース5、7、18を介してトランザクション及び照会プロセッサからのページ読出し及び書込み要求を扱う版マネージャ13から成る。知能ページ・メモリ内の不揮発性メモリは、トランザクション・データベース・ログ6及びトランザクション・データベース・データ8に対する貯蔵所として機能する。このトランザクション・データベース・データは、図1において8として示されているものと全く同一である。つまり、これは、トランザクションプロセッサ6が不揮発性メモリに書込んだ個々のページの現在のコピーに対するパッキング・メモリである。知能ページ・メモリは照会バージョン・データ14のページのための追加の不揮発性メモリを提供する。これらは、データベースの一貫した照会ビューをインターフェース18内の要求を介して照会プロセッサに与えることを可能にする。

【0028】版マネージャ13の機能は、トランザクションプロセッサ及び照会プロセッサによるデータの共有論理ページへのアクセスを制御する一方、これらが互いの性能に大きな影響を与えることを阻止し、また、物理不揮発性メモリに対する総要件を低減することにある。この結果として、トランザクション・データベース・データ8及び照会バージョン・データを保存するために必要とされる総不揮発性メモリは、先行技術においてトランザクション・データベース・データに対して必要とされた量の二分の一以下となる。

【0029】版マネージャ13は、照会プロセッサからのページ読出し要求に応答して、インターフェース18

を介してデータベース・データの最近の一貫したバージョンを提供し、また、同時に、インターフェース5及び7を介してのトランザクションプロセッサからの要求に応答して単純な不揮発性メモリのように振る舞う。

【0030】長時間実行複合照会が照会によって読出されたデータのトランザクション更新をロックアウトすることを阻止する一方において一貫したアクセスを保持するために、照会は、データの一貫した照会スナップ・ショットを見る。つまり、トランザクション更新は、論理的に分離されたトランザクションバージョンにされる。ただし、同一ページの殆どをデータのトランザクションと照会版の間で共有するために、全ての対応するトランザクション及び照会バージョンの論理ページは、その照会バージョンがデータベース・スナップ・ショットと呼ばれるプロセスによって生成されて以来トランザクションによって更新されたページに対する物理メモリの単一ページからサポートされる。我々は、独立したトランザクション及び照会ビューをサポートするために共有物理ページを使用するこの方法を暗黙バージョニング（implicit versioning）と呼ぶ。この方法は図3及び4を使用してより詳細に説明される。

【0031】知能ページ・メモリはまた新たなタイム・ステップがいつ取られるかを決定するための機構を含む。これは、内部アルゴリズムあるいはユーザからの外部プロンプト、トランザクションプロセッサあるいは照会プロセッサに基づく。含蓄バージョニングを実現するために、バージョン・マネージャ13はトランザクション及び照会処理の両方からのページ・アクセスを正しい物理ページにルーティングし、ページのバージョンを保持及び管理し、新たな照会スナップ・ショットの生成を開始及び処理し、また古いページのバージョンから物理メモリを回復及び再使用する責務を持つ。

【0032】知能ページ・メモリ10内の不揮発性メモリ12上に保存されたトランザクション・データベース・ログ6は、トランザクションプロセッサ3がデータベース・データをトランザクション破棄及びシステム故障から保護するために不揮発性メモリに保存するために必要とする正にその情報である。知能ページ・メモリ内のバージョン・マネージャ13は、ログ・インターフェース5内の読出し及び書込み要求に応答して、知能ページ・メモリ不揮発性メモリ上のトランザクション・データベース・ログ6内にこの情報を保存し、また、これをその後の読出し要求に対してリターンする。知能ページ・メモリ内にデータベース・ログを持つことの長所は、このログ情報をトランザクションプロセッサの邪魔をすることなく、あるいはこのスループットを低減することなく、データベース・データの一貫したバージョンを生成するために使用できることである。ログは更新トランザクションに対してのみ保持され、照会プロセッサはトランザクション・ログ情報へのアクセスを必要としないために、ページ・メモリ内にデータベース・ログを格納するために要求される不揮発性メモリの量は、先行技術によるようにログがトランザクションプロセッサに直接に接続された場合と同一である。

【0033】同様に、バージョン・マネジャーは、インターフェース7内のトランザクション・データベース・データ・ページ読出し及び書込み要求に応答して、ページ・イメージをトランザクション・データベース・データ8に対する知能ページ・メモリ内の不揮発性メモリ内に保存し、これらをその後の読出し要求に対してリターンする。これは、バージョン・マネージャがデータベース・スナップ・ショット処理によってその照会バージョンが生成されて以来あるトランザクションがそのページを修正したときにのみ照会に対するデータベースの一貫したビューをサポートするために照会バージョン・データ14内にページの追加の物理コピーを生成することを可能にする。不揮発性メモリは、データベース設備のコストの大きな要素となるために、トランザクション及び照会プロセッサに対するページの不要な別個の物理コピーを回避することは重要である。図3及び図4において説明される暗黙バージョニングには、照会に対する一貫したビューを提供することの要件、及び不揮発性媒体のトランザクション処理に対するアピアランスを満たすためにデータベース・データ・ページのコピーをいつ生成すべきかを決定するための効率的な体系を含む。これは、照会及びトランザクションがデータベース・ページの不要な重複なしに、従って最低限のコストで同時に実行できるようにする。

【0034】知能ページ・メモリ10内の不揮発性メモリ12は、トランザクション・データベース・ログ6、トランザクション・データベース・データ8及び照会バージョン・データを格納するための任意の標準の不揮発性媒体（例えば、磁気ディスク）を使用して実現することができる。

【0035】「暗黙バージョニング：照会バージョンの管理」図3は、暗黙バージョニングを定義する状態図であり、データベース状態、照会スナップ・ショット、トランザクション及び照会間の時間の進行に従っての関係を示す。この状態図は、照会及び照会バージョンが典型的なトランザクションの寿命の100倍あるいは1000倍以上の寿命を持つと言う意味において完全なスケールを持つものではない。我々は、現データ及び一つの照会スナップ・ショットへのトランザクション・アクセスがサポートされるケースに対する暗黙バージョニングについて説明する。この体系の率直な拡張は、新たなスナップ・ショットが生成されるときの崩壊を低減し、追加の不揮発性メモリを犠牲にして追加の照会バージョンを許す。

【0036】図3は、状態・時間図であるが、時間は、左から右へと進行する。互いのすぐ上の事象は、同時に

13

発生するものである。

【００３７】図３において、最も上のセクションは、それぞれトランザクションプロセッサ上で実行されるサンプル更新トランザクションＸ１、Ｘ２、．．．Ｘ７の寿命を示す。これらボックスの各々の左端は、更新トランザクションがいつ実行を開始し、またロック及びデータベース資源の保持がいつ開始されるかを示す。右側は、これがいつ終了し、データベース内で更新がコミットされるかを示す。これらトランザクションは重複し（同時実行され）、必ずしもそれらの開始の順番ではない順番にてコミットされることに注意する。

【００３８】図３の次のセクションは、更新トランザクションＸ１、Ｘ２、．．．Ｘ７が加えられた結果としてのデータベースの状態を時間とともに示す。初期データベース状態２６は、開始時間におけるデータベースの状態を表わす。データベース２７、２８、２９、３０、３１、３２、３３のその後の状態は、トランザクションＸ１、Ｘ２、．．．Ｘ７が順番にコミットした後の変化した状態を示す。トランザクション・データ・ベース状態は、各々が一つのトランザクションの更新を含む小さな粒度のステップにて進行することに注意する。トランザクションがＸ１をＸ２の前に、Ｘ２をＸ３の前にコミットするこの順番は、トランザクションプロセッサ３によって生成されるトランザクション・データベース・ログ６内に正確に反映される。これら状態図内においては、先行するトランザクションからの全ての更新が含まれ、その後のトランザクションからの更新は含まれない。例えば、状態３０、つまり”トランザクションＸ４の後のデータベース状態”は、Ｘ１、Ｘ２、Ｘ３及びＸ４からの更新を含むが、Ｘ５、Ｘ６あるいはＸ７からの更新は含まない。我々は一貫したと言う用語をデータベース状態のこの特性を定義するために使用する。

【００３９】図３の３４、３５、３６、３７を含む次のセクションはデータベース状態のスナップ・ショットを取ることによって生成され、照会によって使用されるために保持される照会バージョンを示す。

【００４０】動作３４において、バージョン・マネージャは３６によって示される寿命を持つ照会バージョンＶ０を生成するために初期データベース状態２６のスナップ・ショットを取る。３６の長さは、照会プロセッサ上で実行される照会によって使用することが可能な照会バージョンＶ０の寿命を示す。動作３５は、トランザクションＸ３がコミットした後のある時間におけるバージョン・マネージャを示す。このコミットは、一貫したデータベース状態２９のスナップ・ショットを取り、これを使用して照会バージョンＶ１を生成することによって行なわれる。照会に対して照会バージョンＶ１が使用できる寿命は、３７の長さによって示される。

【００４１】図３の次のセクションは、サンプルとしての照会寿命を示す。つまり、Ｑ１及びＱ２が処理されて

14

いる期間が、それぞれ、アクティブ寿命３８及び３９として示される。この時間を通じて、Ｑ１及びＱ２は、”照会バージョンＶ０”３６のデータへのアクセスを持たなければならない。照会Ｑ３及びＱ４は、それぞれ、寿命４０、４１を持ち、これら照会は、”照会バージョンＶ１”３７のデータへのアクセスを持たなければならない。

【００４２】バージョン・マネージャ１１によって実現されるこの暗黙バージョニング・体系は、トランザクションプロセッサ３がデータのページをインターフェース７を介して、そのバッファ４の使用を最適化するような方法にて書き出すと言う事実に対処しなければならない。より具体的には、ページは、コミットされないデータを含めて、書き出されたセットのデータがデータベース状態２６、２７、２８、２９、３０、３１、３２、３３の意味において、トランザクション・データベースの一貫した状態を表わすと言う保証なしに書き出される。知能ページ・メモリは、これらページを受信し、照会プロセッサ内のより長い実行期間を持つ照会によって見られるデータのビューを妨害することなく、これらをその後の読出し要求に対してリターンしなければならない。つまり、これらは照会バージョン３６、３７を見なければならない。暗黙バージョニングにおいては、これは、照会によって見られるデータのバージョンを離散ステップにより進めることによって達成される。個々のタイム・ステップにおいて、新たな照会スナップ・ショットが生成され、照会に対して見えるようにされる。個々のスナップ・ショットは、ある最近の時間におけるコミットされたデータ・ベース・データの一貫したビューである。時間ステップの間において、知能ページ・メモリは、照会に対してデータの一貫したビューを与える。トランザクションプロセッサが更新されたページを知能ページ・メモリに書き出すとき、これら更新されたページは、保存されるが、ただし、次の時間ステップまで照会に対して見えるようにされない。

【００４３】「暗黙バージョニング：ページ・コピーの管理」暗黙バージョニングはまた現在のトランザクション及び照会ビューをサポートするためにデータベース・データ・ページの追加のコピーがいつ要求されるかを正確に決定する。これは、不揮発性メモリ要件を最小限にする。暗黙バージョニングによるページ・コピーの管理が図４の時間状態図によって記述される。時間は、この図においては、左から右へと進む。互いにすぐ上の事象は、同時に発生する。

【００４４】知能ページ・メモリ１０内にデータベース・データの初期状態、つまり、トランザクション・データベース・データ８と照会バージョン・データ１４が組み合わされたものが４７として示される。これは、実際には、セットのページ値を格納するセットの論理ページＰ１、Ｐ２、．．．Ｐ７によって表わされる複合状態で

15

ある。我々は、これらページ内に格納される値を、それぞれ、文字”ａ”、”ｂ”、”ｃ”、”ｄ”、”ｅ”、”ｆ”、”ｇ”によって表わす。このデータベース状態は、トランザクション処理が休止した後に再開したばかりのときに起こるように、トランザクション・データベースの一貫した状態及びトランザクション・データの実際の状態であると想定される。図３の状態２６は、このケースを示す。我々は、この一貫した状態のスナップ・ショットが動作３４に示されるように取られ、”照会バージョンＶ０”３６のように、これに対応する照会バージョンが生成されるものと想定する。

【００４５】鍵となる点は、この状態４７において、どのページも二度は格納されないことである。つまり、ページＰ１、Ｐ２、．．．Ｐ７の個々の単一コピーのみでデータベース・データの正しいトランザクション及び照会ビューをサポートするのに十分である。このセットの物理ページは、トランザクション・データベース・データ８として機能する。つまり、この時点において、照会バージョン・データ１４に対して追加の物理メモリは要求されない。

【００４６】トランザクションプロセッサからのインターフェース７を介しての一連のページ読出し及び書込み要求動作４２、４３、４４、４５は、知能ページ・メモリ内に格納されたページ・データに影響を与える。動作４６は、データベースのサブショットを取ることによって新しい照会バージョンをその後生成する効果を示す。これら動作の結果としてのデータ・メモリの状態のシーケンスが状態４８、４９、５０、５１、５２として示される。

【００４７】動作４２は、トランザクションプロセッサからのページＰ５の値を読出すと言う要求である。これは、ページ・メモリ内の”ｅ”という現在の値を受信する。ページＰ１、Ｐ２、．．．Ｐ７に対して格納された値は、修正されておらず、コピーは作成されない。

【００４８】次の動作４３は、トランザクションプロセッサからの値”ｘ”をページＰ３に書込むと言う要求である。状態４９は、古い値”ｃ”及び新しい値”ｘ”の両方が保存できるようにページＰ３のコピーが作成されることを示す。この時点において、照会は、照会バージョン・データ１４内に在駐する古い値”ｃ”を見、一方、不揮発性メモリへのトランザクションプロセッサ要求は、トランザクション・データベース・データ内のページのコピー内の新しい値”ｘ”を見る。

【００４９】次の動作４４は、トランザクションプロセッサからのページＰ３を読出す要求である。これは、前にトランザクションプロセッサの要求によってそこに書込まれたページの値”ｘ”を受信する。動作４４の後のページ・メモリの状態である状態５０は、ページ・メモリの状態がこの読出し動作の結果として修正されなかったことを示す。

16

【００５０】動作４５は、トランザクションプロセッサの値”ｙ”を書込む要求の後のページＰ３へのその後の書込みである。状態５１は、新たな値”ｙ”が古いトランザクション値”ｘ”上に書込まれたが、ただし、この時点において照会によって必要とされる値、つまり、”ｃ”が既にコピー内に保存されているためにページの新たなコピーは作成されないことを示す。

【００５１】動作４６において、データベースのスナップ・ショットを取ることによって新たな照会バージョンが生成されると、ページＰ３に対する古い照会値”ｃ”は破棄することができ、状態５２によって示されるように、このコピーに対して使用された不揮発性メモリが再使用のために回復される。値”ｙ”をページＰ３内に書込む責務を持つトランザクションが”スナップ・ショット”が作られる前にコミットされたと想定すると、これが次の時間ステップにおいて照会に対して提供される値となる。

【００５２】第一の照会期間内における照会プロセッサからの読出し要求は、この時間期間内のいつこれらの要求が行なわれたかに係らず”照会プロセッサからのＶ０読出し”５３によって示されるようにページＰ１、Ｐ２、．．．Ｐ７に対して、値”ａ”、”ｂ”、”ｃ”、”ｄ”、”ｅ”、”ｆ”、”ｇ”を見る。次の期間からの読出し要求は、”照会プロセッサからのＶ１読出し”によって示されるように、論理ページＰ１、Ｐ２、．．．Ｐ７に対してこれら新たな値”ａ”、”ｂ”、”ｙ”、”ｄ”、”ｅ”、”ｆ”、”ｇ”を見る。

【００５３】状態５２までには、照会バージョン・データ１４のために使用された物理メモリは、次の照会バージョンによって再使用されるよう回復されることに注意する。

【００５４】図４は論理的なものである。つまり、現在のデータベースは、一例としての記述において使用される７ではなく、通常、１０，０００以上のデータ・ページを含む。

【００５５】「暗黙バージョニングを実現するバージョン・マネージャ内の論理」図５は、暗黙バージョニングを実現するために使用されるバージョン・マネージャ１３内の論理を示す。これは、制御フローのグラフである。バージョン・マネージャ内のデータ構造は、以下から成る。

・ファイル・マップに対するページ

これは、その論理ページの主コピーが格納されているファイル・システム内の位置に対するページ番号をマップする。トランザクション・データベース状態が現照会バージョンと同一のときは（例えば、図４内の状態４７）、トランザクション・データベース・データ８は、データベース内の各々の論理ページの主コピーから成る。

17

・コピー標識ベクトル
これは、データベース内の個々の論理ページに対する1ビットの情報を含み、照会バージョン・データ14の一部として副ピーが作成されているか否かを示す。
・コピー・インデックス
これは、作業メモリ内の副ページの位置を示す。これは、例えば、B－Treeの形式を持つ。

【0056】次に、図5の流れ図の説明に移る。バージョン・マネジャーへの事象入力55は、トランザクションプロセッサからの"照会プロセッサからページを読出す"（58）ための、あるいは"新たな照会バージョンを生成する"（59）ための"ページを読出すための要求"（56）、あるいは"ページの書込み"（57）である。

【0057】事象56が受信されると、つまり、読出すページの番号及び期待されるリターン・ページの値とともに、"トランザクションプロセッサからのページを読出す要求"が受信されると、動作60は、ファイル・マップに対するページを使用してこのページの位置を調べる。動作61は、このページの値をファイル・システムのこの位置から読出すことによって得て、動作62は、要求に応答してこの値をリターンする。

【0058】事象57が受信されると、つまり、書込むページの数及び書込むべき値とともに"トランザクションプロセッサからのページ書込み"が受信されると、動作63によって、そのページの副コピーが現照会バージョンによって必要とされるデータにて既に作成されているか否かを決定するためにそのコピー標識ベクトルがチェックされる。まだ作成されてないときは、動作64がページ・フレームの作業スペースから照会バージョン・コピーに対するスペースを割り当て、主コピー内に前に格納されているそのページの照会値をこうして割り当てられたスペースにコピーし、コピー・インデックスをこのコピーを指すように更新し、この論理ページに対応するコピー標識ベクトル内の標識をコピーが既に作成されていることを示すようにセットする。

【0059】このポイント以降は、処理は、照会コピーの作成を必要としたか否かに関係なく同一である。動作65において、主コピーの位置がファイル・マップに対するページを使用して要求されたページ番号を検索することによって決定され、この新たな値が動作66において前の値の上に書込まれ、この要求が次に動作67によって完結される。

【0060】我々の上記の好ましい実施態様においては、照会バージョンが"アウト・オブ・ポジション（out of position）"にされ、最も最近のトランザクション値に対して主コピーが使用されると言う照会及びトランザクション・システムの特性を持つ。ただし、この選択は、暗黙バージョニング概念に本質的なものではない。

18

【0061】事象58が受信されると、つまり、読出すべきページの番号及び期待されるリターン値とともに"照会プロセッサからのページ読出し要求"が受信されると、動作68において、このページの照会コピーが既に作成されているか否かを決定するためにコピー・ベクトルがチェックされる。既に作成されているときは、動作72においてコピー・インデックスを使用してこのコピーの位置が決定され、動作73においてこの位置から値が読出される。既に作成されてないときは、動作69においてファイル・マップに対するページを使用して主コピーの位置が決定され、動作70において、ファイル・システム内のこの位置から読出すことによってページ値が得られる。動作71は、要求者にこれらの経路のいずれかによって得られた値をリターンする。

【0062】新たな照会バージョンを作成したい場合、事象59において、セットの主ページがトランザクション・データベースの一貫した状態を表わす場合、副ページが開放され、これらの内容が動作74において破棄される。次に、動作75において、コピー・インデックス及びコピー標識ベクトルが副照会バージョンのページが現在存在しないことを示すようにリセットされる。最後に、動作76において、この事象に対する処理が完結したことが示される。

【0063】ページ・メモリ内のセットの主ページがトランザクション・ベースの一貫した状態を表わすことを保証するための方法が以下に説明される。

【0064】「ページ・メモリに一貫したトランザクション・データベース状態をもたらすための方法」ディスクにデータベースのトランザクションの一貫したスナップ・ショットをもたらすための一つの方法は、トランザクションプロセッサ上の全てのトランザクションを、これらを休止させ、つまり、これらがデータベースを出た時点で直ちにこれらを停止し、バッファ4内に入れることによって、一貫した状態に保つ方法である。

【0065】トランザクション処理主体が修正されコミットされたページがメモリ内にディスクにフラッシュされることなく留まることができる時間の長さに制限を与えるような場合は、（トランザクション処理に対して最少の妨害を持つ）好ましい実施例は、以下に説明されるトランザクション・トラヒックに妨害を与えないタイム・ステップのためのアルゴリズムである。

【0066】「保証されたページ・アウトを持つタイム・ステップの詳細」このセクションにおいては、任意の修正されたページが時間Tの前に（これがバッファ内において修正され、トランザクションがコミットされた後に）、書き出されることを保証するデータベースに対するタイム・ステップ・アルゴリズムについて詳細に説明する。・時間tに対して、A（t）をその時間tにおいてアクティブである全てのトランザクションの開始時間の最少であると定義する。A（t）は、常に、tの前で

19

あるかｔに等しい。・ＬＳＮ　ｔが時間ｔにおいて到達されるログ・オフセットであるものとする。Ａ（ｔ）よりも前の全てのトランザクション活動は、時間ｔまでに、ＬＳＮ　ｓの所のログ内のページ更新に対して解決されているために（ここで、ＬＳＮ　ｓ＆Ａ１．ＬＳＮ　Ａ（ｔ）、ログ期間（ＬＡＭ　ｓ、ＬＳＮ　ｔ）を調べることによってこの更新がコミットされたか破棄されたかを決定することが可能である。・時間Ｔ内のページ・アウトを保証するためのトランザクション処理主体とは、何を意味するかと言う厳密な定義を与える。つまり、時間ｓにおいてコミットされる任意のページ更新に対して、この更新（及び恐らくは後の動作）を含むページのバージョンは、時間ｓ＋Ｔまでにディスクにフラッシュされ；時間ｓにおいて破棄され、デッスクにページの正しくないバージョンがフラッシュされることを許される任意のページ更新に対しては、実行されなかった更新（及び恐らくはオーバーライトされた後の動作）を持つページのバージョンが時間ｓ＋Ｔまでにディスクにフラッシュされる。・時間Ｔ、及び任意の与えられた時間ＴＳ内のページ・アウトを保証するトランザクション処理主体に対しては、Ａ（ＴＳ－Ｔ）より前の全てのページ更新は、（責務を持つトランザクションが時間ＴＳ－Ｔまでにコミットあるいは破棄し、動作の結果が少なくとも時間Ｔの後にディスクにフラッシュされることが保証されるために）時間ＴＳまでに”ディスク上に”反映される。・時間ＴＳにおいて、照会処理時間ステップを取るためには、以下の動作が必要である。

【００６７】１．時間ＴＳの直後に、書込むためのトランザクション処理の全て動作は、時間ＴＳにおいて存在したイメージについてページの予備照会コピーを作成することを強要する。

【００６８】２．時間Ａ（ＴＳ）及びＡ（ＴＳ－Ｔ）が決定される。これは（知能ページ・メモリ内において）、ログ内のチェック・ポイント・レコード内に書き出されたトランザクション・テーブルから開始し、新たなトランザクションの開始及び終端を示しながらログを通じて前方向に実行することによって遂行される。

【００６９】３．時間ＴＳにおける知能ページ・メモリ内のデータ・ベースのイメージに対してログ期間（ＬＳＮ　Ａ（ＴＳ－Ｔ）、ＬＳＮ　ＴＳ）を処理することによって、時間Ａ（ＴＳ）までにコミットされた全てのトランザクションを反映するデータ・ベースの正しいスナップ・ショットを構成することが可能である。

【００７０】４．個々の全ての更新手順Ａ（ＴＳ－Ｔ）がコミットあるいは破棄され、この更新の対応する効果あるいは破棄（undo）が知能ページ・メモリ内の予備照会イメージ内に反映されることに注意する。

【００７１】５．ログを通じてのＬＳＮ　Ａ（ＴＳ－Ｔ）からＬＳＮ　ＴＳへの順方向のパスが行なわれる。ＬＳＮ　Ａ（ＴＳ）の前の個々のページ更新に対して、

20

ＬＳＮＡ（ＴＳ）前に予定となっているトランザクションがコミットされたか否かを決定することができる。つまり、期間（ＬＳＮ　Ａ（ＴＳ）、ＬＳＮ　ＴＳ）内のページがＡ（ＴＳ）までにコミットを終えることはない。さらに、更新レコードのログ位置を予備照会ページ内のＬＳＮと比較することによって、我々は、Ａ（ＴＳ）の前にコミットする任意の更新がそこに反映されているか否かを決定することができる。我々は、このような更新を順番に（ログを通じて前方向に）満たすために、これらがページの予備照会コピー内に存在しないときは、これらを加えることができる。ログからのページ更新を処理するためのアルゴリズムは以下の通りである。

【００７２】（ログを通じて順方向に進む）個々のページ更新に対して、

ｉｆ　トランザクションがＡ（ＴＳ）前にコミットする時はｉｆ　ＬＳＮ（更新）＆Ａｒ．ＬＳＮ（予備ＱＤＢページ）であるときは、ｔｈｅｎ　更新を行なう。

－－これは、我々がログ内のＬＳＮ　Ａ（ＴＳ－Ｔ）から開始したためである。

－－これは、”シーケンス内の次の更新”でなければならない。

ｅｌｓｅ　ＬＳＮ（更新）以下、あるいはＬＳＮ（予備ＱＤＢページ）以上であるときは、スキップする。

－－この更新は予備ＱＤＢコピー内に既に反映されている。

ｅｌｓｅ　Ａ（ＴＳ）の後にトランザクションが破棄あるいはコミットされている。

（このカテゴリーにおいては、全ての更新は、Ａ（ＴＳ）の後である）。

ｉｆ　ＬＳＮ（更新）＆Ａｒ．ＬＳＮ（予備ＱＤＢページ）の場合は、ｔｈｅｎスキップする。

－－更新は、予備ＱＤＢにフラッシュ・アウトされてない。

ｅｌｓｅ　ＬＳＮ（更新）＝ＬＳＮ（予備ＱＤＢページ）

この更新、及びそのページに対する、前にスタックされている全てのアンドウをアンドウ（破棄）する。

－－この更新は、ページ・メモリに対して行なわれるが、その後は行なわーれない。実行されたページのバージョン；更新を無効し、その後そのーーページに対して前にスタックされている全ての更新を逆の順番に無効ーーにする。

ｅｌｓｅ　ＬＳＮ（更新）＆Ａ１．ＬＳＮ（予備ＱＤＢページ）の場合、全てのアンドウ（破棄）が逆の順番に存在する予備ＱＤＢに対して適用できる場合、後に適用するためにそのアンドウ（破棄）を積み重ねる。

【００７３】６．このログ・パスが完了した時点で、予備ＱＤＢページは、Ａ（ＴＳ）によってコミットされた全てのトランザクションに対するデータ・ベース・イメ

21

ージを反映する。

・メモリ内にページの積み重ねられたアンドウ（無効）動作を管理するための空きが存在しないときは、これは、ログ・レコードにマークを付け、これらを連結し、ログを通じてのＬＳＴ　ＴＳからＬＳＮ　Ａ（ＴＳ－Ｔ）に向かっての第二の逆方向へのパスを行なうことによって処理することができる。

・ログ処理の際に予備ＱＤＢページが修正される度に、これは、”停止及びフラッシュ（Suspend and Flush）”アルゴリズムにおいて記述されるページのイメージを保存する技術を使用して遂行される。形式的には、

ｉｆ　予備ＱＤＢコピーが既に存在する場合は、

－－ページのＡＤＢバージョンから分離し、

ｔｈｅｎ　現存の予備ＱＤＢバージョンを修正する。

ｅｌｓｅ

－－時間ＴＳイメージにて予備ＱＤＢバージョンを生成し、これに、アンドウあるいは再実行（redo）動作を適用する。

・全てにおいてページ・アウト保証時間が同一であるべきであると言う理由は存在しない。時間ステップに対して時間ＴＳが選択された後、唯一必要とされることは、その時間の前の全てのページ更新の影響がＴＳまでに解決されディスクにページ・アウトされるように時間を決定することである。メモリ内で修正されているが、ページ・アウトされてないページの汚れたページ・テーブル（Dirty Page Table、ＤＰＴ）を維持するデータ・ベースに対しては、Ａ（min（ＤＰＴ（ＴＳ））がことの特性を持つ。

・時間ステップの説明にあたって、トランザクション処理主体の故障及び再開に関してはまだ説明してない。最も単純で一般的なアプローチ（そして、我々の好ましい実施例）は、トランザクション処理主体が故障した場合、これを次の照会時間ステップが取られる前に回復する方法である。時間ステップ・アドバンスとトランザクション・システムの回復とを織り混ぜる方法は、可能ではあるが、トランザクション・システムによって使用されるある特定の回復戦略においては問題がある。

【００７４】ページ・アウトの保証は、トランザクション処理システムの故障が発生しないという前提の下で、全ての解決された更新がその時間内にディスクにフラッシュされる時間である。

【００７５】「データベース及びログ・メモリの所有権」上の説明において、我々は、知能ページ・メモリがデータベース・ページ及びトランザクション処理からのログに対して使用されるメモリの所有権を持つものと想定した。

【００７６】Ａ（ｔ）のような量の計算は、知能ページ・メモリがログ内のチェック・ポイントからアクティブのトランザクション・テーブルを読み、このログからその後のトランザクションの開始及び終端動作を知るようにすることによって最も簡単に実現できる。

22

【００７７】同様に、Ｔを計算する最も簡単な方法は、ログ・チェック・ポイント・レコードから汚れたページ・テーブルを読出し、一層の精度が要求される場合は、これを時間毎に前進させ、その後ラッシュされたページが存在するかメモリをチェックする方法である。

【００７８】「時間ステップ前進の際に照会処理への妨害を防ぐ方法」前述のように、全ての照会は、データの正当な照会バージョンに対して実行されなければならない。従って、新たなスナップ・ショットが生成されている最中に照会が実行されてはならない。照会処理のこの中断は、個々の修正されたページの最大一つの追加のページを機性にすることによって回避することができる。こうすることによって、最後の照会バージョン（Ａ）に対して実行される照会が継続でき、この間に、新たな照会が新たな照会バージョン（Ｂ）を見ることができる。さらにもう一つの照会バージョン（Ｃ）は、全ての照会アクセス・バージョン（Ａ）が完結し、バージョン（Ａ）が削除されるまで生成されない。

【００７９】「トランザクションと照会プロセッサ間のインデックス及びメタデータの共有」知能ページ・メモリが関連テーブル・データのバージョニングとの関連で説明された。同様の技法が修正を必要とすることなくトランザクションプロセッサ・カタログ及び他のメタデータによって構築されたインデックスへのアクセスを照会プロセッサに与えるために使用できる。バージョニングは、このように照会プロセッサに対して使用できるようにされたインデックス及びメタデータ・ページが照会テーブル・データと完全に同期され、一貫性を持つことを保証する。本発明が好ましい実施態様の形式で説明されたが、本発明の真の範囲及び精神から逸脱することなく、形式あるいは上の説明の細部について修正することが可能であることは勿論である。

【００８０】

【発明の効果】本発明によれば、共有不揮発性メモリを含む知能ページメモリに対してデータベース・トランザクション処理及び照会処理を同時に行うようにすることができる。

【図面の簡単な説明】

【図１】先行技術によるトランザクション及び照会システムの関連するシステム構造を示す略図である。

【図２】本発明の好ましい実施態様に対するシステム構造の略図であり、トランザクション及び照会によるデータへの同時及び一貫したアクセスのための知能ページ・メモリが示される。

【図３】データベース・スナップショットがいつ生成されるか、これらに対して一貫とは何を意味するか、及び照会バージョンを保持するとは何を意味するかを示す状態／時間図である。

【図４】ページ・メモリ内の各々のページのコピーの数

23

が時間を通じてトランザクションプロセッサからの書込み要求及び新たなバージョンの生成の結果としていかに変動するかを示す状態／時間図である。

【図５】好ましい実施態様において暗黙バージョニングを実現するために知能ページ・メモリによって使用されるアルゴリズム及び論理を定義する図である。

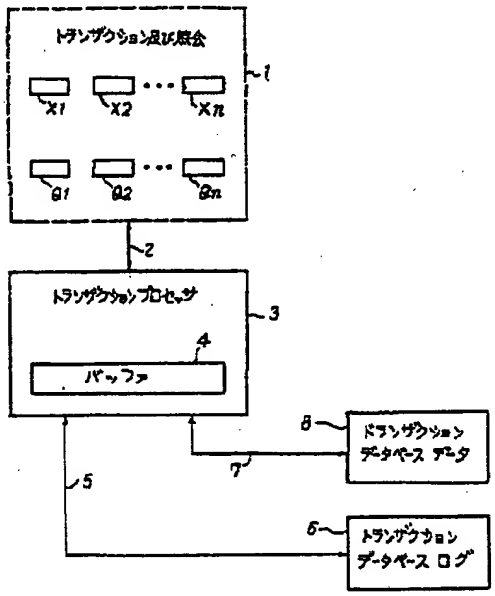
【符号の説明】

3　トランザクションプロセッサ
4　バッファ
6　トランザクション・データベース・ログ

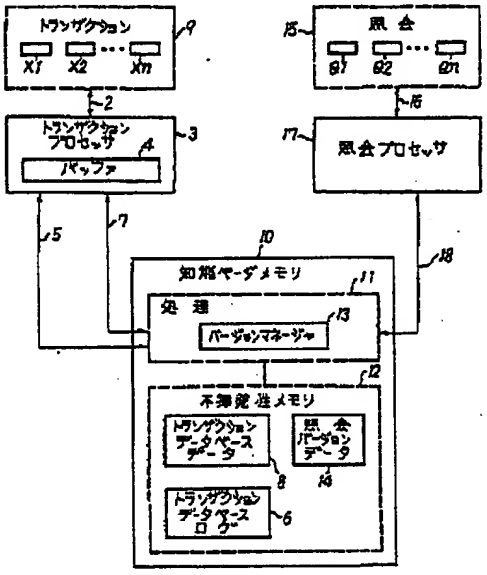
24

8　トランザクション・データベース・データ
9　トランザクション
１０　知能ページメモリ
１１　処理
１２　不揮発性メモリ
１３　バージョンマネージャ
１４　照会バージョンデータ
１５　照会
１７　照会プロセッサ

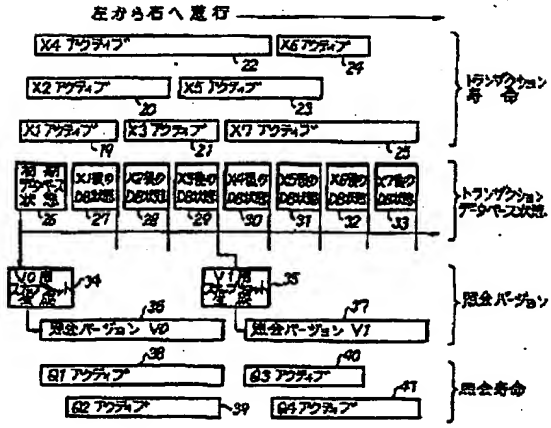
【図１】



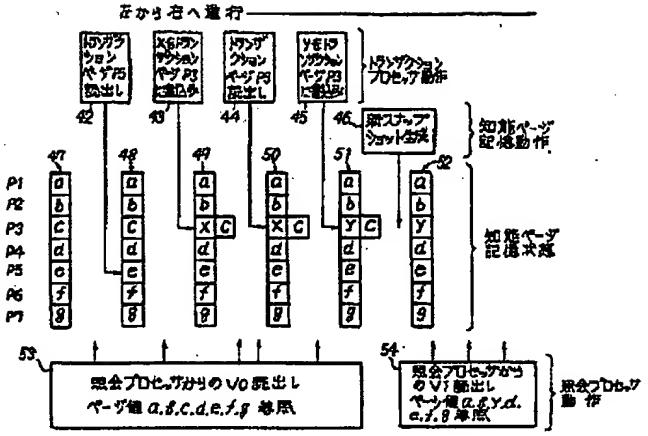
【図２】



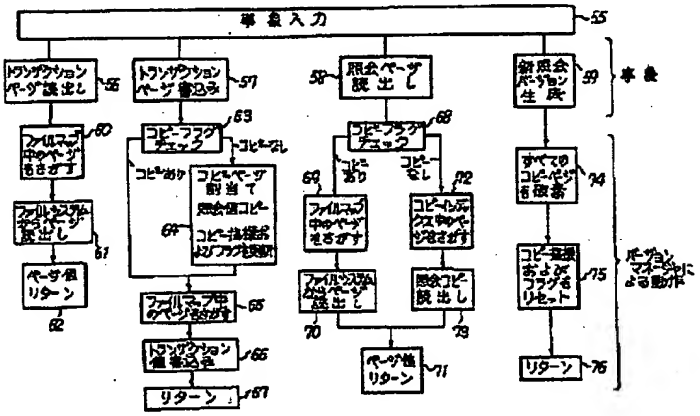
【図３】



【図4】



【図5】



フロントページの続き

(72)発明者　アムブジ、ゴヤール
アメリカ合衆国ニューヨーク州、アマウオーク、ノエル、コート、ボックス、172

(72)発明者　フランシス、ニコラス、バー
アメリカ合衆国ニューヨーク州、クロトン－オン－ハドソン、ナンバー、632、アールエフデイー、1、テイータウン、ロード、82

Japanese Unexamined Patent Publication No. 6-314227
Published on November 8, 1994

Title of the Invention

A LOCKING MECHANISM FOR VERSION OBJECT

Abstract

[Object] To provide a method and an apparatus for controlling simultaneous access to a version object by a plurality of users in a distributed data processing environment.

[Construction] Firstly, there is retrieved a writing lock for a first portion of a data set to which a user desires an access. When the writing lock is not detected, the user's request is permitted by allowing writing access to the first portion. On the contrary, when the writing lock is set, another user is permitted to get a reading access to the first portion although he is inhibited to change the writing access. When a second user desires an access to a second portion of the data set, and the second portion has several elements common to the first portion and unshared elements, the request by the second user is permitted partially for a part of the second portion unshared by the first portion.